Behavior Modification

http://bmo.sagepub.com

Identifying Evidence-Based Interventions for Children and Adolescents Using the Range of Possible Changes Model: A Meta-Analytic Illustration

Andres De Los Reyes and Alan E. Kazdin

Behav Modif 2009; 33; 583 originally published online Aug 14, 2009;
DOI: 10.1177/0145445509343203

The online version of this article can be found at:
http://bmo.sagepub.com/cgi/content/abstract/33/5/583

Published by:
SAGE
http://www.sagepublications.com

Behavior Modification
Volume 33 Number 5
September 2009 583-617

10.1177/0145445509343203
http://bmo.sagepub.com
hosted at
http://online.sagepub.com

Identifying Evidence-Based Interventions for Children and Adolescents Using the Range of Possible Changes Model

A Meta-Analytic Illustration

Andres De Los Reyes
University of Maryland at College Park
Alan E. Kazdin
Yale University

The article discusses a study involving a framework (range of possible changes [RPC] Model) developed and applied to identify patterns in consistent and inconsistent intervention outcomes effects by informant, measurement method, and method of statistical analysis to the meta-analytic study of trials testing two evidence-based interventions for children and adolescents (youth-focused cognitive-behavioral treatment for child anxiety problems; parent-focused behavioral parent training for childhood conduct problems). This article illustrates how findings gleaned from applying the RPC Model allow for unique opportunities for hypothesis generation based on the patterns of consistent outcomes effects. Based on the RPC Model, studies can be closely examined to identify the specific instances in which interventions yield robust effects, and the authors illustrate how examining effects in this way can lead to new understandings of interventions and the outcomes they produce. Findings suggest that researchers can employ previously underutilized

Authors' Note: This work was supported, in part, by National Institute of Mental Health Grant MH67540 (Andres De Los Reyes). This work was also supported by National Institute of Mental Health Grant MH59029 (Alan E. Kazdin). We are very grateful to Kelly D. Brownell, Julia Kim-Cohen, Susan Nolen-Hoeksema, and Peter Salovey for extremely insightful discussions and commentaries on previous versions of this manuscript. We also thank Jennifer Thomas, Jessica Cronce, and Amelia Aldao for their careful and diligent participation as coders for this study. Please address correspondence to Andres De Los Reyes, Department of Psychology, University of Maryland at College Park, Biology/Psychology Building, Room 3123H, College Park, MD 20742; office: 301-405-7049; e-mail: adelosreyes@psyc.umd.edu.

patterns of consistencies and inconsistencies in outcomes effects as new resources for identifying evidence-based interventions.

Keywords: *efficacy; effectiveness; intervention, range of possible changes*

One of the most important research questions addressed in the applied sciences is whether the interventions developed to target specific problem domains or behaviors (e.g., anxiety, delinquency, hypertension, mood) produce meaningful change. A key goal of this research is to identify treatment techniques supported by experimental evidence or evidence-based interventions (EBIs). Intervention research is a key factor of joint efforts from policy makers at the local, state, and federal level, as well as of academic institutions that develop lists of recommended interventions for specific conditions (e.g., American Psychological Association Interdivisional Task Force on Child and Adolescent Mental Health, 2007). Therefore, a critical need exists to develop reliable and valid approaches to identifying EBIs.

To identify a single intervention as an EBI, researchers conduct numerous studies of that intervention and have developed approaches for classifying or quantifying evidence accumulated over numerous studies. Currently, the two most often employed approaches are (a) categorical classification systems; and (b) meta-analytic research reviews. With these approaches, researchers identify EBIs as those interventions that successfully target the constructs they were developed to target (e.g., Roth & Fonagy, 2005; Weisz, McCarty, & Valeri, 2006). However, recent work suggests these approaches do not address the inconsistent evidence often gathered within and between intervention studies (De Los Reyes & Kazdin, 2008). For instance, current approaches do not take into account the consistency by which individual outcomes findings within and between studies yield similar conclusions, despite the absence of definitive methods for distinguishing between "right" and "wrong" findings (De Los Reyes & Kazdin, 2006).

The common presence of inconsistent outcomes evidence within controlled trials raises fundamental questions about how to review and classify evidence from intervention research and how this evidence ought to be employed to identify EBIs. Indeed, inconsistent findings can arise across outcomes measures depending on the informant, measurement method, or method of statistical analysis (De Los Reyes & Kazdin, 2006). For instance, discrepancies commonly arise across multiple informants' ratings of the same construct or set of behaviors, such as reports of symptoms of anxiety and mood, pain, and parenting taken from such informants

as patients, clinicians, laboratory observers, and parents and teachers in the case of children (Achenbach, McConaughy, & Howell, 1987; Achenbach, Krukowski, Dumenci, & Ivanova, 2005). These discrepancies have been identified as one of the most robust findings identified in the clinical sciences (Achenbach, 2006; De Los Reyes & Kazdin, 2005). In particular, they have implications for the treatment of children and adolescents because multiple informants are commonly employed in intervention research with this patient population (Weisz, Jensen Doss, & Hawley, 2005). Given the common presence of inconsistent outcomes, it is important to determine whether this variability in outcomes evidence yields meaningful information in its own right. In particular, might these inconsistencies be employed to further understand intervention effects and determine the circumstances in which interventions work best?

Recently, a framework was developed to take into account inconsistent findings and identify important circumstances in which interventions yield robust effects. Broadly, the Range of Possible Changes (RPC) Model takes into account inconsistent findings by operationally defining consistency and inconsistency in the identification of significant intervention effects, across multiple-outcome measures of the same outcome domain (e.g., symptom and diagnostic presentation; De Los Reyes & Kazdin, 2006, 2008). The RPC Model does this in two ways. First, the framework includes classification categories that researchers may employ to identify consistencies in significant effects within studies and between studies of the same intervention (Table 1). Second, the RPC Model focuses on identifying ranges of the magnitudes of differences between conditions (e.g., ranges of effect sizes across outcome measures) so that researchers may identify patterns of effects across categorical and dimensional outcome approaches (Table 2). Therefore, applying the RPC Model to a systematic review of controlled outcome studies may advance the literature on reviewing outcomes evidence because this would demonstrate why outcomes should be observed within ranges of findings tapping into the same domain of intervention outcomes. Furthermore, a key strength of the RPC Model is that it combines elements of both categorical classification criteria (Table 1) and quantitative review approaches (effect size ranges; see Table 2) to identify EBIs. The RPC Model has been applied to conceptualizing treatment outcomes for specific conditions, such as social skills training for childhood pervasive developmental disorders (Koenig, De Los Reyes, Cicchetti, Scahill, & Klin, 2009). However, the RPC Model has yet to be applied across multiple studies so that it would be possible to statistically compare variability in outcomes effects both within and between studies.

Table 1
Description and Criteria of RPC Model Categories[a]

Category	Criteria
Best evidence for change	At least 80% of the findings from three or more informants, measures, and analytic methods show differences, and at least three findings were gleaned from each of the informants, measures, and methods. There is no clear informant-specific, measure-specific, or method-specific pattern of findings. The evidence suggests the intervention successfully targets the construct.
Evidence for probable change	More than 50% of the findings from three or more informants, measures, and analytic methods show differences, and at least three findings were gleaned from each of the informants, measures, and methods. There is no clear informant-specific, measure-specific, or method-specific pattern of findings. The evidence suggests the intervention probably changes the targeted outcome domain, yet future work ought to examine why inconsistencies occurred.
Limited evidence for change	Either 50% or less of the findings from three or more informants, measures, and analytic methods show differences, or less than the grand majority (less than 80%) of findings from specific informant's ratings, measures, and/or methods show differences. Any differences found are either scattered across outcomes from multiple informants, measures, or methods, or are not found predominantly on outcomes from specific informants, measures, and/or methods. The evidence is inconclusive.
No evidence for change	No differences are observed. The evidence is completely inconclusive.
Evidence for informant-specific change	Differences are found on the grand majority (80%) of ratings provided by specific informant(s), and at least three findings were gleaned from the informant(s) for which specificity of findings were observed. The evidence suggests the treatment might change the domain when it is exhibited in specific situations or in interactions with specific informant(s).
Evidence for measure- or method-specific change	Differences are found on the grand majority (80%) of specific measure(s) or analytic method(s), and at least three findings were gleaned from the measure(s) or method(s) for which specificity of findings were observed. The evidence suggests the intervention might change the domain when it is measured with specific kinds of measure(s), method(s), or both.

Note: In the categories above, by "informants" we mean reporters of outcomes (e.g., self, spouse or significant other, clinician, laboratory observer, biological, institutional records); by "measures" we mean ways to assess outcomes (e.g., questionnaire or symptom-count measures, laboratory observations, diagnostic interviews); by "analytic methods" we mean statistical strategies (e.g., tests of mean differences, tests of diagnostic status). For a discussion on the development and rationale for the structure and criteria of the individual RPC Model categories, see De Los Reyes and Kazdin (2006).

a. Adapted from De Los Reyes and Kazdin (2006) and De Los Reyes and Kazdin (2008).

Table 2
Description and Criteria of Effect Size Ranges to be Employed in Conjunction With Range of Possible Changes Model Categories

Category	Criteria
Below small to below small	Lower end includes any effect size below .20. Upper end includes any effect size below .20.
Below small to small	Lower end includes any effect size below .20. Upper end includes any effect size greater than or equal to .20, but less than .50.
Below small to medium	Lower end includes any effect size below .20. Upper end includes any effect size greater than or equal to .50, but less than .80.
Below small to large	Lower end includes any effect size below .20. Upper end includes any effect size greater than or equal to .80.
Small to small	Lower end includes any effect size greater than or equal to .20, but less than .50. Upper end includes any effect size greater than or equal to .20, but less than .50.
Small to medium	Lower end includes any effect size greater than or equal to .20, but less than .50. Upper end includes any effect size greater than or equal to .50, but less than .80.
Small to large	Lower end includes any effect size greater than or equal to .20, but less than .50. Upper end includes any effect size greater than or equal to .80.
Medium to medium	Lower end includes any effect size greater than or equal to .50, but less than .80. Upper end includes any effect size greater than or equal to .50, but less than .80.
Medium to large	Lower end includes any effect size greater than or equal to .50, but less than .80. Upper end includes any effect size greater than or equal to .80.
Large to large	Lower end includes any effect size greater than or equal to .80. Upper end includes any effect size greater than or equal to .80.

Note: Effect size ranges based on Cohen's *d* (1988).

This article identifies key issues in reviewing the evidence for EBIs by applying the RPC Model to a meta-analytic review of two EBIs (youth-focused cognitive-behavioral treatment for child anxiety problems; parent-focused behavioral parent training for child conduct problems). We examine multiple studies of these two treatments that themselves serve as exemplary controlled trials in that they assessed outcomes of the construct targeted for treatment using multiple informants, measures, and methods of statistical analysis. We will illustrate how an RPC Model meta-analysis identifies patterns in outcomes evidence gleaned within and between studies that

assessed outcomes in multiple ways. Specifically, we will demonstrate that the use of the RPC Model will have the critical utility of identifying specific instances in which interventions produce consistent effects. Furthermore, we provide recommendations for research on the identification of EBIs. In doing so, we discuss the conceptual and research implications of applying the RPC Model in future meta-analytic and experimental research.

Method

Interventions Examined

The sample examined in the meta-analysis consisted of a set of elegantly conducted randomized controlled clinical trials that tested the efficacies of two specific interventions, each developed to target a specific psychological construct: (1) youth-focused cognitive-behavioral therapy for childhood anxiety problems (hereafter referred to as CBT), and (2) parent-focused behavioral parent training for childhood conduct problems (hereafter referred to as BPT).[1] We defined CBT using the definition employed by Weisz, Hawley, and Jensen Doss (2004): an individual, youth-focused intervention entailing "efforts to identify and alter cognitions that contribute to the anxiety and to identify and alter maladaptive behavior (such as avoidance of feared situations) that may serve to sustain the condition" (p. 751). Similarly, we defined BPT using a definition employed by Weisz et al. (2004): those interventions focused on parents that aim to reduce child conduct problems by employing some or all of the following components:

(1) Parents learn basic behavioral principles relevant to child rearing;
(2) Parents learn how to define, track, and record rates of the antisocial and prosocial behaviors they want to target;
(3) Parents are helped to design, role-play, carry out, and refine behavior modification programs while continuing to record rates of target behavior to assess intervention effects. (p. 792)

Literature Review

The literature review and collection of studies were accomplished on two fronts. First, studies published in years up to and including 2002 were derived from a previous meta-analysis of the youth intervention literature (Weisz et al., 2004). Second, literature searches for relevant intervention studies published between the years 2003 through 2006 were conducted

employing the same methods as Weisz et al. (2004). Two standard computerized databases were employed to identify relevant studies. We used Psychinfo, limiting our search from 2003 to 2006, and employing 21 psychotherapy-related keywords derived from prior meta-analytic work (see Weisz et al., 1987, 1995). We also conducted searches of the same years using MEDLINE, via PubMed. We limited our search from 2003 through 2006, and used the same search terms as Weisz et al. (2004): *Mental Disorders* with the search limits: *clinical trial*, *child* (*3-18 years*), *published in English*, and *human subjects*.

Criteria for Study Inclusion and Study List

Consistent with prior work, we focused the review on examinations of peer-reviewed intervention outcome studies (see Lonigan et al., 1998; Nathan & Gorman, 2007; Roth & Fonagy, 2005; Weisz et al., 1987, 1995, 2004). Among these peer-reviewed studies, we employed stringent criteria for study inclusion, based on prior work (Weisz et al., 2004). Specifically, studies included in the review were required to meet the following criteria: (1) The intervention being examined must have been compared to an inert control group, such as waitlist, no treatment, placebo, or other inert process; (2) each study must have employed a prospective design and random assignment of participants to conditions; (3) each study must have examined a sample of youths within a 3- to 18-year-old age range; (4) each study must have examined participants selected for exhibiting the behavior or emotional problems identified previously (child anxiety, child conduct); (5) each study must have employed a postintervention assessment of the construct being targeted for intervention; (6) participants in groups being compared to one another must not have been taking psychotropic medications; (7) each study must have reported in their analyses a statistically significant benefit of the intervention examined, relative to a control condition on at least one outcome measure of the target construct; and (8) each study must have employed at least three measures of the construct targeted for intervention (i.e., three anxiety measures for studies of CBT, three conduct problem measures for studies of BPT).[2]

Study list. The study list included 9 studies of CBT and 7 studies of BPT, yielding 21 intervention-control comparisons (11 for CBT, 10 for BPT). Lists of the studies and descriptions of basic demographic, methodological, and outcome measure characteristics are presented in Tables 3 and 4. Although the number of studies on the list is less than the total number of controlled

(text continues on p. 594)

Table 3
Demographic Characteristics of CBT and BPT Studies Included in the Meta-Analysis

Study	Type of Sample	Pretreated Sample Size	Age Range of Total Sample	% Boys	Pretreated Sample Size of Experimental Conditions, Meta-Analysis	Number of Intervention Groups, Meta-Analysis	Number of Intervention Groups, Total Sample
CBT studies							
Barrett et al. (1996)	Diagnosed outpatients, clinic referred	79	7-14	56.96	54	1	2
Flannery-Schroeder & Kendall (2000)	Diagnosed outpatients, clinic referred	45	8-14	51.51	45	2	2
Gallagher et al. (2004)	Diagnosed outpatients, recruited sample	23	8-11	47.83	23	1	1
Kendall (1994)	Diagnosed outpatients, clinic referred	47	9-13	60.00	47	1	1
Kendall et al. (1997)	Diagnosed outpatients, clinic referred	118	9-13	62.00	118	1	1
King et al. (2000)	Symptomatic outpatients, clinic referred	36	5-17	30.56	24	1	2
Leal et al. (1981)	Symptomatic school sample	30	10th grade students	N/A[a]	30	2	2
McMurray et al. (1986)	Symptomatic school sample	80	9-12	50.00	80	1	1
Spence et al. (2000)	Diagnosed outpatients, clinic referred	50	7-14	62.00	33	1	2

(continued)

Table 3 (continued)

Study	Type of Sample	Pretreated Sample Size	Age Range of Total Sample	% Boys	Pretreated Sample Size of Experimental Conditions, Meta-Analysis	Number of Intervention Groups, Meta-Analysis	Number of Intervention Groups, Total Sample
BPT studies							
Leung et al. (2003)	Symptomatic outpatients, clinic referred	91	3-7	63.77	91	1	1
Webster-Stratton (1984)	Symptomatic outpatients, clinic referred	35	3-8	71.43	24	1	2
Webster-Stratton et al. (1988)	Symptomatic outpatients, clinic referred	114	3-8	69.30	114	3	3
Webster-Stratton (1990)	Symptomatic outpatients, clinic referred	47	3-8	79.07	47	2	2
Webster-Stratton (1992)	Symptomatic outpatients, clinic referred	100	3-8	72.00	100	1	1
Webster-Stratton & Hammond (1997)	Diagnosed outpatients, clinic referred	97	4-8	74.23	48	1	3
Webster-Stratton et al. (2004)	Diagnosed outpatients, clinic referred	159	4-8	90.00	57	1	5

CBT = cognitive behavioral treatment; BPT = behavioral parent training.
a. Leal et al. (1981) did not provide this information.

Table 4
Methodological and Outcome Characteristics of CBT and BPT Studies Included in the Meta-Analysis

Study	Intervention-Control Comparison	Number/ Informants	Number/ Measure Methods	Number/ Analytic Methods	Number/ Outcome Measures	Number Outcome Findings	Statistically Significant Findings (%)	Pretreatment Difference
CBT studies								
Barrett et al. (1996)	ICBT versus WL	3	2	2	4	5	2 (40.00)	No
Flannery-Schroeder & Kendall (2000)	ICBT versus WL	3	1	1	6	6	4 (67.00)	No
	GCBT versus WL	3	1	1	6	6	6 (100.00)	Yes[a]
Gallagher et al. (2004)	GCBT versus WL	3	2	2	5	6	3 (50.00)	No
Kendall (1994)	ICBT versus WL	3	2	2	6	7	5 (71.43)	Yes[b]
Kendall et al. (1997)	ICBT versus WL	3	1	1	5	6	5 (83.33)	Yes[c]
King et al. (2000)	ICBT versus WL	3	2	1	4	7	3 (42.86)	No
Leal et al. (1981)	GCBT (CM) versus WL	2	2	1	3	3	0	No
	GCBT (SD) versus WL	2	2	1	3	3	0	No
McMurray et al. (1986)	GCBT versus Placebo	2	2	1	3	3	1 (33.33)	N/A[d]
Spence et al. (2000)	GCBT versus WL	4	3	1	5	5	1 (20.00)	No
BPT studies								
Leung et al. (2003)	GBPT versus WL	1	1	1	3	4	4 (100.00)	Yes[e]
Webster-Stratton (1984)	GBPT versus WL	2	2	1	3	5	3 (60.00)	No
Webster-Stratton et al. (1988)	GBPT (VM) versus WL	3	2	1	4	8	7 (87.50)	No
	GBPT (GD) versus WL	3	2	1	4	8	4 (50.00)	Yes[f]
	IBPT versus WL	3	2	1	4	8	5 (62.50)	No

(continued)

Table 4 (continued)

Study	Intervention-Control Comparison	Number/ Informants	Number/ Measure Methods	Number/ Analytic Methods	Number/ Outcome Measures	Number Outcome Findings	Statistically Significant Findings (%)	Pretreatment Difference
Webster-Stratton (1990)	IBPT (VM) versus WL	3	2	1	4	5	1 (20.00)	No
	IBPT (VM/TC) versus WL	3	2	1	4	5	0	No
Webster-Stratton (1992)	IBPT versus WL	3	2	1	4	7	5 (71.43)	No
Webster-Stratton & Hammond (1997)	GBPT versus WL	3	2	1	5	7	5 (71.43)	Yes[g]
Webster-Stratton et al. (2004)	GBPT versus WL	4	2	1	3	3	2 (66.67)	No

CBT = cognitive behavioral treatment; BPT = behavioral parent training; ICBT = individual CBT; GCBT; Group CBT; CM = cognitive modification; SD = systematic desensitization; IBPT = individual BPT; GBPT = group BPT; WL = waitlist; VM = video modeling; GD = group discussion; TC = therapist consultation.

a. Three measures were significant between conditions preintervention.

b. Two measures were significant between conditions preintervention.

c. One measure was significant between conditions preintervention.

d. The authors reported employing outcome measures prior to intervention to identify anxious youths to participate in the study, but did not report preintervention scores. However, the authors did not report significant preintervention differences between conditions.

e. One measure was significant between conditions preintervention.

f. One measure was significant between conditions preintervention.

g. One measure was significant between conditions preintervention.

outcome studies of CBT and BPT within the date range of our review, this number is consistent with prior reviews of carefully selected studies of cognitive-behavioral treatments of childhood anxiety, as well as parenting and family treatments for child behavior problems (James, Soler, & Weatherall, 2005; Woolfenden, Williams, & Peat, 2001).

Study Coding Procedures

Coding manual. A coding manual was developed to describe procedures for coding information from studies (manual available from the authors). Briefly, the manual was separated into multiple parts and developed to outline and describe coding procedures for basic study characteristics (e.g., sample size, type, and demographics; Part 1), outcome measure characteristics (information source, outcome measure methodology; Part 2), effect size and statistical test calculations (Part 3), and classifications of studies based on the RPC Model (Part 4). Mean effect size calculations were made using statistical software.

Coding descriptions and reliability. Three clinical science graduate students were trained to code information gleaned from studies. All three coders were blind to the study hypotheses. Two coders were trained to individually code all information in each of the 16 studies. One coder with experience in conducting and coding information for meta-analyses was trained as a consensus coder with the key tasks of leading coding meetings. Each of the 16 studies was separately coded in their entirety by the two coders, and the consensus coder led reliability meetings with both coders present. Specifically, in these meetings, the consensus coder led discussions of each item coded for each study, led discussions on resolutions of coding inconsistencies between the two coders, and recorded the number of instances in which inconsistencies were evident between the two coders. Resolution of coding inconsistencies was reached by consensus from all coders, and the consensus coder recorded a final code in these circumstances. Across items coded within the 16 studies in the meta-analysis, the consensus coder resolved coder inconsistencies 3.5% of the time for Parts 1 to 4 (156 out of 4,481 items). The rate of inconsistencies within each section was as follows: Part 1: 7.1% (64 out of 904 items), Part 2: 4.5% (45 out of 1,005 items), Part 3: 1.4% (30 out of 2,138 items), and Part 4: 3.9% (17 out of 434 items).

Coder training. In order to ensure that reliability of codes gathered in one section did not influence the level of reliability of codes gathered from

other sections, coders first were trained and coded information concerning basic study and outcome measure characteristics and statistical test and effect size calculations. Once this information was coded, consensus codes were distributed to all coders for use in coding information for the 16 studies, for codes relevant to classifying the evidence for individual studies.

Coder training for basic study and outcome measure characteristics and statistical test and effect size calculations was accomplished by practicing applying the coding manual to seven studies that were excluded from the list of studies coded in the meta-analysis. These excluded studies included studies reporting controlled trials of treatments for CBT and BPT. Coding practices relevant to classifying evidence under the RPC Model involved having coders evaluate and code results of 14 hypothetical studies. Coding for the 16 studies commenced after the study coders agreed that the coding manual and sheets were sufficiently clear, and enough experience was accrued in practices for coders to report feeling adequately prepared.

Postintervention effect size calculations. Calculations of effect sizes were performed for each of the methods employed by studies to examine intervention outcomes (mean differences, diagnostic status, clinically significant change, see Cohen, 1988; Cohen, Cohen, West, & Aiken, 2003; Rosenthal & DiMatteo, 2001). Mean differences calculations were made by subtracting the control group mean from the intervention group mean, and dividing this difference by the control group's standard deviation at outcome (Glass's Δ; see Rosenthal & DiMatteo, 2001). Glass's Δ is an effect size metric that meta-analysts consider being within the *d* family metric of effect sizes (Rosenthal & DiMatteo, 2001). Thus, we maintained a consistent presentation of effects across methods of analysis by presenting Glass's Δ results using the *d* symbol. Diagnostic status and clinically significant change calculations were calculated using the Phi (ϕ) coefficient to examine differences in proportions between conditions (see Cohen et al., 2003). There were instances in which only results of statistical tests where available (e.g., *t* statistic). Thus, effect sizes (using the *r* metric) were estimated in these instances using test statistics, as suggested elsewhere (Rosenthal & DiMatteo, 2001). Given the use of *r* effect size measures for some calculations and that ϕ is an *r* effect size measure as well, effect sizes calculated using ϕ and *r* were converted to *d*, in order to construct effect size ranges along a common metric (Rosenthal & DiMatteo, 2001). Lastly, all effect sizes were adjusted to take into account small sample bias, employing Hedges' small sample correction (Hedges & Olkin, 1985).

Calculating and coding statistical tests of intervention outcomes. In addition to calculating effect sizes, we evaluated the consistency of statistical tests of differences between intervention and control conditions. Studies were quite variable in the methods employed to examine statistical differences between conditions. However, the statistical power of significance tests is influenced by sample size and the type of statistical test employed (Cohen, 1988). Thus, calculations of statistical differences between conditions were kept constant across examinations of statistical test outcomes within and between studies. Specifically, for tests of mean differences, coders recorded the posttreatment means and standard deviations for each intervention and control condition and for each outcome measure, and employed an online independent samples *t*-test calculator to code the results (Graphpad Software, Inc., 2005).

For tests of diagnostic status and clinically significant change, coders recorded the postintervention frequencies of participants in intervention and control conditions for each dichotomous outcome measure and employed an online chi-square test statistical calculator to code the results of significant tests of diagnostic status and clinically significant change (Ball, 2003). All statistical tests were conducted as two-tailed tests to take into account the possibility of both positive and negative effects, and the threshold for statistical significance was set at $p < .05$. Sometimes studies reported an intervention outcome using only the result of a statistical test such as *t* test or chi-square. In these instances, coders recorded the statistical information and employed that information to both calculate effect sizes and code statistical significance.

Preintervention group comparability. Finally, as an added check on results of postintervention analyses as well as the fidelity of random assignment in each study, coders recorded preintervention information for each of the measures, using procedures identical to codes of postintervention results.

RPC Model: Incorporating study codes, effect sizes, and intervention effects. Coders employed the RPC Model to classify the findings gleaned from each of the 21 intervention-control comparisons conducted across the 16 studies. To make RPC Model categorical classifications coders reviewed information they coded previously on outcome measure and statistical outcome characteristics (outcome measure methodology, outcome measure source, method of statistical analysis), along with the results of statistical outcomes (statistical tests, effect size calculations). Coders were provided with a copy of the original table from De Los Reyes and Kazdin (2006) denoting criteria for the RPC Model categories. The RPC Model's origin was not disclosed to coders during data collection.

Coders made RPC Model classifications for each intervention-control comparison (see Table 1). Specifically, for each intervention-control comparison, coders identified the percentages of findings that were statistically significant, based on information coded previously. These percentages of findings were employed to determine the RPC Model category within which a given study would be classified. Furthermore, if specificity in significant effects could be identified (e.g., three or more findings based on parent report yielding consistently significant effects and could be classified in the "evidence for informant-specific change" category) and findings could not be classified in a nonspecific effect category (e.g., "evidence for probable change"), the study was classified in an RPC Model category denoting specificity in intervention effects (e.g., "evidence for informant-specific change"). For the purposes of these RPC Model classifications as well as for all other calculations and classifications, we defined "finding" as any single instance in which an intervention-control comparison was made on an outcome measure of the construct targeted for intervention. Under such a definition, a single outcome measure could contribute more than one finding if (a) the measure was examined using more than one statistical method, and/or (b) more than one subscale within that measure was examined using one or more statistical methods. Study classifications using the RPC Model as well as calculations of mean effect sizes were based on the nature and extent of these findings. For a discussion on the development and rationale for the structure and criteria of the individual RPC Model categories, see De Los Reyes and Kazdin (2006).

Furthermore, each intervention-control comparison was coded for the range of effect sizes (i.e., upper and lower limit effect sizes) observed within its RPC Model classification. Coders coded these effect size ranges for findings within the RPC Model category classifications for each of the 21 intervention-control comparisons. For each intervention-control comparison and its RPC Model category classification(s), coders identified the highest and lowest observed effect sizes. Under all circumstances, effect size ranges consisted of all findings employed to reach the RPC Model category classification, irrespective of whether the finding was statistically significant or if the effect size was negative (i.e., intervention condition had worse scores, relative to controls). In addition, coders employed Cohen's (1988) effect size conventions of small (.20), medium (.50), and large (.80) effects to categorize effect size ranges. Furthermore, coders were instructed to consider any effect sizes below .20 (including negative effect sizes, where the intervention had worse scores, relative to controls) under a new category: below small. The criteria employed to construct these ranges are presented in Table 2, and the ranges captured every possible effect size

limit that could be reached, based on Cohen's (1988) conventions. Coders employed this system to categorize these ranges.[3]

Quantifying mean effect sizes. We were interested in comparing effect size findings gleaned from the RPC Model to the mean effect size gleaned across studies. Given that study inclusion criteria pertaining to number of outcome measures employed resulted in each study providing more than one effect size, effect sizes were aggregated within the study so that a mean effect size could be attained for each intervention-control comparison. In addition, in cases in which an RPC Model classification of a study was for specificity in change (Table 1), this might result in the inclusion of some effect sizes gleaned from the study included in effect size ranges and others left out. Therefore, in order to provide as conservative a test as possible, all effect sizes gleaned from the study were employed to calculate mean effect sizes for the study, regardless of the RPC Model classification for that study. A further consideration is that the RPC Model was developed to classify individual intervention-control comparisons. Thus, studies examining more than one intervention yielded more than one data point. Therefore, we calculated and reported a mean effect size for each intervention-control comparison and categorized this mean effect size as a below small, small, medium, or large effect, consistent with Cohen's (1988) effect size conventions (Table 2).

Data analytic plan. The main analyses compared the mean upper and lower limit effect size findings identified within studies both to each other and to estimates of mean effect sizes across studies. These comparisons were addressed in two ways. First, we conducted a paired-samples *t* test to compare the mean lower limit effect size and the mean upper limit effect size gleaned from individual intervention-control comparisons. Second, we conducted two one-sample *t* tests, one comparing mean lower limit effect sizes across intervention-control comparisons with the mean effect size across intervention-control comparisons, and another comparing mean upper limit effect sizes across intervention-control comparisons with the mean effect size across intervention-control comparisons. Results reported below were consistent, regardless of whether one-sample or paired-samples *t* tests were employed. Furthermore, we ran the same analyses comparing RPC Model upper and lower limit effect sizes with the mean intervention-control comparison effect size, excluding the upper and lower limit effect sizes for each intervention-control comparison from calculations of the mean. Results from these analyses were consistent with results reported below, suggesting that outlier effects could not explain our findings.

All analyses were conducted within the entire sample for a specific intervention (i.e., separate for CBT and BPT studies). In addition, hypotheses were directional and tested with low statistical power, given the small number of studies examined in the meta-analysis. Therefore, all tests comparing the RPC Model to other approaches were conducted as one-tailed significance tests, and effect sizes for analyses were calculated using Cohen's *d* based on the test statistics yielded from statistical analyses.

Results

Basic Study Characteristics

Basic study characteristics for CBT and BPT studies are presented in Table 3. For CBT studies, the mean preintervention study sample was 50.44 (*SD* = 30.96), based on the size of treatment and control conditions that were studied in the meta-analysis: control, *M*(*SD*) = 19.50 (13.50); intervention, *M*(SD) = 21.80 (17.84); eight studies provided this data. The overall sample percentage mean attrition was 13.10% (*SD* = 9.59%; seven studies provided this data). Youths in CBT studies tended to range in age from middle childhood to adolescence. Participants tended to include slightly more numbers of boys (*M* = 52.61, *SD* = 10.47; eight studies provided this data). Mean treatment length was 803.41 minutes (*SD* = 424.92).

For BPT, the mean preintervention study sample was 68.71 (*SD* = 33.07); control, *M*(*SD*) = 26.86 (12.75); and intervention, *M*(*SD*) = 29.30 (14.03). The overall sample percentage mean attrition (including conditions not meeting inclusion criteria) was 9.27% (*SD* = 8.56%; six studies provided this data). Youths in BPT studies tended to range in age from preschool to early childhood. Furthermore, participants tended to be boys (*M* = 74.26; *SD* = 8.35). Mean treatment length was 1173.40 minutes (*SD* = 840.64).

Methodological and Outcome Measure Characteristics

Methodological and outcome measure characteristics for CBT and BPT studies are presented in Table 4. For both CBT and BPT studies, all data reported in this section were with reference to individual intervention-control comparisons. For CBT, more than one data point came from two studies. All but one of the intervention-control comparisons compared the intervention of interest to a waitlist control condition; the other intervention-control comparison employed a psychological placebo comparison condition.

The CBT intervention-control comparisons employed a mean number of outcome measures of 4.55 (*SD* = 1.21) and a mean number of outcome findings derived from those measures of 5.18 (*SD* = 1.54). The CBT comparisons employed an average of 2.82 informants (*SD* = .60), 1.27 outcome analytic methods (*SD* = .47), 1.18 types of statistical tests (*SD* = .40), and 1.82 outcome measure methodologies (*SD* = .60). Overall, CBT intervention-control comparisons yielded a mean of 2.73 findings that were statistically significant (*SD* = 2.10), which translated to a mean overall percentage of significant findings of 46.18% (*SD* = 32.45). However, these percentages of significant findings were not necessarily the ones employed to make RPC Model classifications (e.g., RPC Model classifications of informant, measure, and/or method specificity might yield different percentages; see Table 1).

For BPT studies, more than one data point came from two studies. All studies compared BPT interventions to a waitlist control condition. The BPT intervention-control comparisons employed a mean number of outcome measures of 3.80 (*SD* = .63) and a mean number of outcome findings derived from those measures of 6.00 (*SD* = 1.83). The BPT comparisons employed an average of 2.80 informants (*SD* = .79) and 1.90 outcome measure methodologies (*SD* = .32). All 10 BPT comparisons exclusively employed mean differences comparisons as the outcome analytic method and *t* tests as the method of statistical tests. Overall, BPT intervention-control comparisons yielded a mean of 3.60 findings that were statistically significant (*SD* = 2.12), which translated to a mean overall percentage of significant findings of 58.95% (*SD* = 29.73). These percentages of significant findings were not necessarily the ones employed to make RPC Model classifications, particularly those for which specificity in findings are identified (Table 1).

Preintervention Group Comparability

For CBT studies, three studies were identified with preintervention differences on child anxiety measures (Table 4). Specifically, Flannery-Schroeder and Kendall (2000) evidenced significant preintervention differences on three measures, all in the direction of lower intervention group scores relative to control group scores. Kendall (1994) evidenced preintervention differences on two measures, one in the direction of greater intervention group scores and the other in the direction of lower intervention group scores, relative to the control group. Lastly, Kendall et al. (1997) evidenced preintervention differences on one measure, in the direction of

lower intervention group scores relative to control group scores. As reported in the following, each of these studies was identified as suggesting some consistencies in evidence for change under the RPC Model. Therefore, it is unlikely that these differences contributed to fewer opportunities to identify consistencies among measures employed in these studies.

For BPT studies, three studies were identified with preintervention differences between conditions on one child conduct measure (Table 4). Specifically, Leung, Sanders, Leung, and Lau (2003) evidenced preintervention differences on one measure, in the direction of lower intervention group scores, relative to control group scores. Webster-Stratton, Kolpacoff, and Hollinsworth (1988) evidenced preintervention differences on one measure, in the direction of lower intervention group scores, relative to control group scores. Lastly, Webster-Stratton and Hammond (1997) evidenced preintervention differences on one measure, in the direction of greater intervention group scores relative to control group scores. As reported below, two of these studies (Leung et al., 2003; Webster-Stratton & Hammond, 1997) were identified as suggesting consistent findings under the RPC Model. Furthermore, although the third study was not identified by the RPC Model as suggesting consistent findings, the study would have been classified the same, regardless of the results gleaned from the preintervention differences measure or whether that measure was employed at all to classify the study's evidence. Again, it is unlikely that these differences contributed to a decreased likelihood of identifying consistent findings among measures employed in these studies.

Findings Based on the RPC Model

Findings gleaned from the RPC Model yielded substantial variability among study outcomes both in categorical classifications and ranges of magnitudes of intervention effects. Main findings for each study are presented in Table 5 and graphically represented in Figures 1 and 2. Specifically, 3 of the 11 CBT comparisons (28.3%) and 4 of the 10 BPT comparisons (40%) were classified in RPC Model categories suggesting specificities in intervention effects (systematic intervention effects gleaned from specific informants, measures, and/or methods of gauging intervention effects). In two of the three CBT comparisons, consistencies in significant differences between conditions were specific to findings from child self-reported anxiety outcome measures. In three of the four BPT comparisons, consistencies in significant differences between conditions were specific to findings from parent-reported conduct problem outcome measures. Furthermore, as can

(text continues on p. 606)

Table 5
Descriptive Statistics for CBT and BPT Findings Gleaned From the RPC Model and Quantitative Review[a]

Study 1st Author/ Publication Year	Format	Manual 1st Author/ Publication Year	RPC Model Category	RPC Model Lower Limit Effect Size	RPC Model Upper Limit Effect Size	RPC Model Effect Size Range	Mean Effect Size[b]
CBT studies							
Barrett (1996)	Individual	Barrett (1991)	Limited evidence for change	.28	0.65	Small to medium	0.47[c]
Flannery-Schroeder (2000)	Individual Group	Kendall (1990) Flannery-Schroeder (1996)	Limited evidence for change Evidence for informant-specific change Evidence for measure- or method-specific change Specificity: assessed via child self-report, measured via questionnaire, examined via mean differences	.55 .89	2.35 1.54	Medium to large Large to large	1.14[e] 1.26[e]
Gallagher (2004)	Group	Unnamed manual	Limited Evidence for Change	.10	1.13	Below small to large	0.67[d]
Kendall (1994)	Individual	Kendall (1990)	Evidence for Informant-Specific Change Specificity: assessed via child self-report	.40	1.07	Small to large	0.93[e]

(continued)

Table 5 (continued)

Study 1st Author/ Publication Year	Format	Manual 1st Author/ Publication Year	RPC Model Category	RPC Model Lower Limit Effect Size	RPC Model Upper Limit Effect Size	RPC Model Effect Size Range	Mean Effect Size[b]
Kendall (1997)	Individual	Kendall (1990)	Evidence for measure- or method-specific change Specificity: measured via questionnaire, examined via mean differences	.40	.80	Small to large	0.56[d]
King (2000)	Individual	Unnamed manual	Limited evidence for change	.30	1.73	Small to large	0.95[e]
Leal (1981)	Group (CM)	Unnamed manual	No evidence for change	.02	.93	Below small to large	0.44[c]
	Group (SD)	Unnamed manual	No evidence for change	–.47	1.00	Below small to large	0.35[c]
McMurray (1986)	Group	Unnamed manual	Limited evidence for change	.16	.63	Below small to medium	0.34[c]
Spence (2000)	Group	Spence (1995)	Limited evidence for change	.28	1.58	Small to large	0.68[d]
BPT studies							
Leung (2003)	Group	Sanders (1999)	Evidence for informant-specific change Evidence for measure- or method-specific change Specificity: assessed via parent report (unspecified), measured via questionnaire, examined via mean differences	.51	1.06	Medium to large	0.86[e]

(continued)

Table 5 (continued)

Study 1st Author/ Publication Year	Format	Manual 1st Author/ Publication Year	RPC Model Category	RPC Model Lower Limit Effect Size	RPC Model Upper Limit Effect Size	RPC Model Effect Size Range	Mean Effect Size[b]
Webster-Stratton (1984)	Group	Webster-Stratton (1981)	Limited evidence for change	.56	0.88	Medium to large	0.69[d]
Webster-Stratton (1988)	Group (GDVM)	Webster-Stratton (1981, 1987)	Evidence for measure- or method-specific change Specificity: measured via questionnaire, examined via mean differences	.40	1.17	Small to large	0.77[d]
	Individual (IVM)	Webster-Stratton (1981, 1987)	Limited evidence for change	.31	0.98	Small to large	0.57[d]
	Group (GD)	Webster-Stratton (1981, 1987)	Limited evidence for change	.16	1.08	Below small to large	0.60[d]
Webster-Stratton (1990)	Individual (IVM)	Webster-Stratton (1987)	Limited evidence for change	–.13	1.86	Below small to large	0.51[d]
	Individual (IVMC)	Webster-Stratton (1987)	No evidence for change	.41	1.54	Small to large	0.68[d]
Webster-Stratton (1992)	Individual (IVM)	Unnamed Manual	Evidence for informant-specific change Evidence for measure- or method-specific change Specificity: assessed via parent report (mother), measured via questionnaire, examined via mean differences	.51	0.77	Medium to medium	0.56[d]

(continued)

Table 5 (continued)

Study 1st Author/ Publication Year	Format	Manual 1st Author/ Publication Year	RPC Model Category	RPC Model Lower Limit Effect Size	RPC Model Upper Limit Effect Size	RPC Model Effect Size Range	Mean Effect Size[b]
Webster-Stratton (1997)	Group	Webster-Stratton (1990)	Evidence for informant-specific change Evidence for measure- or method-specific change Specificity: assessed via parent report (mother), measured via questionnaire, examined via mean differences	.62	1.47	Medium to large	0.78[d]
Webster-Stratton (2004)	Group	Webster-Stratton (2001)	Limited evidence for change	.39	1.19	Small to large	0.87[e]

CBT = cognitive behavioral treatment; BPT = behavioral parent training; RPC = range of possible changes; CM = cognitive modification; SD = systematic desensitization; GDVM = group discussion videotape modeling; GD = group discussion; IVM = individual videotape modeling; IVMC = individual videotape modeling with therapist consultation.

a. All CBT interventions were compared to a waitlist control condition, except McMurray (1986) (psychological placebo). All BPT interventions were compared to a waitlist control condition.

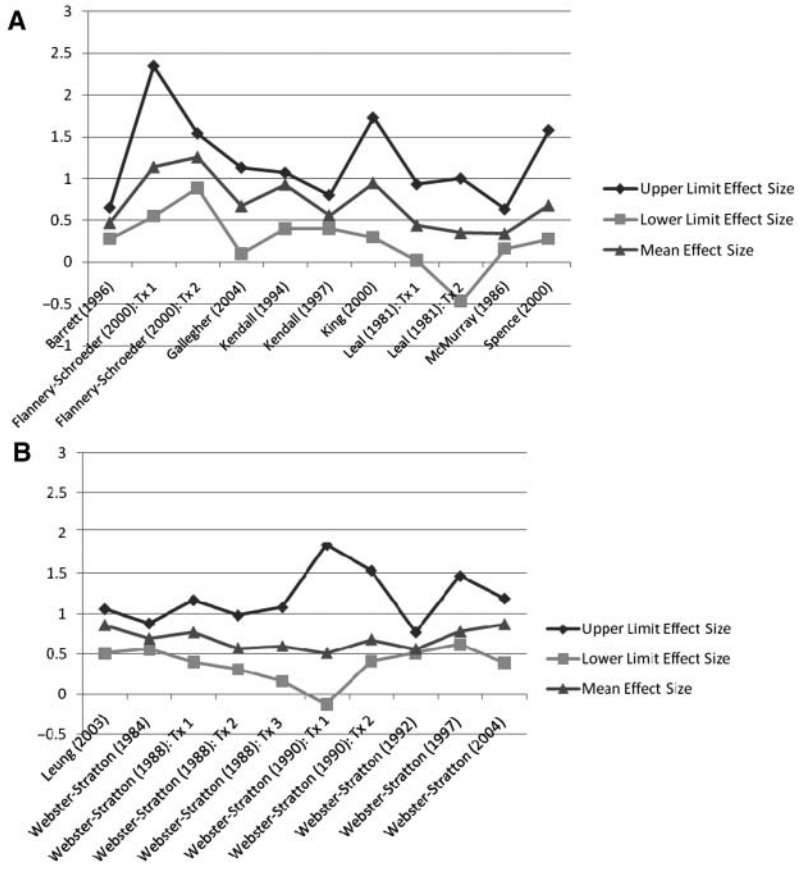
b. Effect sizes in this column were based on all effects gathered in the study, and were classified as small, medium, or large effects, based on criteria outlined by Cohen (1988).

c. Small effects.

d. Medium effects.

e. Large effects.

Figure 1
Upper, Lower, and Mean Effect Sizes for Youth-Focused Cognitive Behavioral Treatments for Childhood Anxiety (top panel, a) and Parent-Focused Behavioral Parent Training for Childhood Conduct Problems (bottom panel, b)



be seen in Figure 1, the variability between upper and lower limit effect sizes was consistent with the variability observed via categorical statistical analyses, even among studies conducted by the same investigator. Indeed, in 7 of the 11 CBT comparisons (63.6%) and 6 of the 10 BPT comparisons

(60%), the evidence suggested that effects ranged from either below small to large (i.e., below .20 to .80 or above) or small to large (i.e., at or above .20 to .80 or above; see Table 5).

Mean Effect Size Comparisons

We identified significant differences between effect size ranges identified via the RPC Model and mean effect sizes (see Table 6). First, paired-samples *t* tests revealed significant differences between RPC Model mean lower limit and upper limit effect sizes gathered within studies. Furthermore, one-sample *t* tests revealed significant differences between each of the RPC Model mean effect size limits and the mean effect sizes gleaned from each set of studies. Thus, results suggest differences between the RPC Model upper and lower limit effects as well as between these two effects and the mean effect size across studies.

Discussion

This article applied a new approach to reviewing and classifying outcomes evidence that takes into account patterns of outcomes findings and identifies the specific circumstances in which EBIs yield consistent effects. The RPC Model was applied to a meta-analysis of a conservative sample of controlled trials of a youth-focused intervention for childhood anxiety problems (CBT) and a parent-focused intervention for child conduct problems (BPT). For both sets of treatments, we identified important differences across outcome measures of the same construct both in whether they suggested categorical differences between intervention and control conditions and in the magnitudes of differences between intervention and control conditions. Most critically, in a number of circumstances these differences varied systematically by how outcome measures were assessed and examined statistically. For instance, for CBT, we identified studies that yielded robust effects specific to child self-report, and for BPT we identified studies that yielded robust effects specific to parent report. Thus, our findings illustrate how support of an intervention can be quite nuanced and dependent on how one evaluates the evidence.

There are limitations to the present study. We can identify two levels of limitations. First, special features of this study might well limit the findings. For example, this study only evaluated within- and between-study consistencies in evidence on a single domain: Diagnostic and symptom

Table 6
Summary Information and Comparisons Between the RPC Model and Quantitative Review for CBT and BPT Studies

Effect Size Range Summary Information	Comparisons Between RPC Model Effect Size Ranges and Means Gleaned From Quantitative Review		
CBT studies			
RPC Model effect size range (*N*, %)	Lower limit effect size mean (*SD*)	Upper limit effect size mean (*SD*)	Comparison effect size mean (*SD*)
Below small to medium: 1 (9.1)%	.26 (.34)	1.22 (.53)	.71 (.32)
Below small to large: 3 (27.3%)	Statistical tests of differences		
Small to large: 4 (36.4%)	Lower limit versus upper limit	Lower limit versus mean	Upper limit versus mean
Small to medium: 1 (9.1%)	–6.47**	–4.36**	3.20*
Medium to large: 1 (9.1%)	Effect size of statistical differences		
	Lower limit versus upper limit	Lower limit versus mean	Upper limit versus mean
Large to large: 1 (9.1%)	–4.13	–2.76	2.02
BPT studies			
RPC Model effect size range (*N*, %)	Lower limit effect size mean (*SD*)	Upper limit effect size mean (*SD*)	Comparison effect size mean (*SD*)
Below small to large: 2 (20%)	.37 (.22)	1.20 (.33)	.69 (.13)
Small to large: 4 (40%)	Statistical tests of differences		
Medium to medium: 1 (10%)	Lower limit versus upper limit	Lower limit versus mean	Upper limit versus mean
Medium to large: 3 (30%)	–5.36*	–4.52*	4.85*
	Effect size of statistical differences		
	Lower limit versus upper limit	Lower limit versus mean	Upper limit versus mean
	–3.53	–2.98	3.23

Note: CBT = cognitive behavioral treatment; BPT = behavioral parent training; Values for *t* tests based on paired-samples tests for comparisons between lower limit and upper limit effects, and one-sample tests comparing each limit effect to the comparison mean effect size for CBT (.71) and BPT (.69) studies.
All tests conducted as one-tailed significance tests, and all effect sizes calculated based on test statistics.
*$p < .01$. **$p < .001$.

presentation of the primary target of treatment. Perhaps findings as to within- and between-study patterns of outcome effects would have been different had the study focused on other domains of functioning. At the same time, focusing on primary target symptom and diagnostic presentation provided as conservative an initial test as possible of the RPC Model in that greater consistencies in evidence might be expected from outcomes assessing the domain targeted for treatment. In addition, focusing on diagnostic and symptom presentation allowed us to identify a number of studies that might not have been identified had the study focused on other domains. Nevertheless, we encourage future research to employ the RPC Model to study within- and between-study consistencies in evidence pertaining to multiple outcome domains.

Furthermore, characteristics of both outcome measures employed (informant, method) and sampling were not held constant across studies. We employed rigorous inclusion criteria to identify studies. This likely constrained variance across studies in design and outcome measure methodology. Nevertheless, between-study consistencies might have been greater had there been even less variability in these characteristics between investigations. However, a number of the studies we reviewed not only tested the same intervention, but the same investigator often conducted multiple tests of the same intervention. Investigators often employed the same outcomes and study designs from study to study. Furthermore, across studies, in many instances the same outcome measures were employed, and in the case of within-study examinations (i.e., sampling characteristics held constant across outcomes), inconsistencies were apparent. In these instances, within- and between-study inconsistencies were apparent, regardless of outcomes assessment and sampling characteristics. In any event, we encourage future research in larger literatures (e.g., controlled trials of psychotropic treatments) to match studies on sampling and outcomes characteristics to ensure that our findings were not due to methodological differences across studies.

The RPC Model also has inherent limitations. To be clear, a model is needed that integrates or interprets the discrepant findings that commonly arise within and between intervention studies. The RPC Model addresses this by identifying patterns by which categorical and dimensional measures of intervention effects yield similar or disparate conclusions. However, other models might well do this. In fact, three limitations of the RPC Model may limit its ability to increase our understanding of inconsistent outcomes effects. First, in examining ranges of effects, use of the RPC Model assumes that the magnitudes of intervention effects (e.g., effect sizes) are of such where large effect sizes indicate more palpable intervention outcomes than

small effect sizes (e.g., $d = .30 < .80$). Indeed, this is a key issue because under some circumstances small effects might signify more salient outcomes than large effects. For instance, one can argue that a small intervention effect on number of police contacts or mortality rates may be far more "important" or "meaningful" than a large intervention effect on mean scores taken from a self-reported behavior checklist. At the same time, the RPC Model's classification categories account for patterns of significant effects specific to particular methods of measuring constructs targeted for intervention. Differential patterns of effects by measurement method might reveal whether studies yield supportive evidence on some measures (objective indices of behavior, official records) and not others (structured interviews, questionnaires). Knowledge of these patterns of effects might address the issues raised by different magnitudes of intervention effects suggested by different forms of measuring behavior. Regardless, future efforts in model development might employ strategies for constructing ranges of effect sizes by weighting effect sizes based on how measurements of behavior were taken (for a further discussion of issues of measurement reliability and other methodological issues see De Los Reyes & Kazdin, 2006).

Second, a limitation of the framework is that it does not address the broader issue of whether interventions make a palpable difference in the lives of clients. This limitation exists, in part because of the arbitrary metrics employed in clinical science to evaluate intervention effects (Blanton & Jaccard, 2006; Kazdin, 2006). For example, a reliable and valid structured interview may be employed to determine whether a depressed client is diagnosis free once the intervention is complete. However, if the measure of presence or absence of a depressive disorder diagnosis is not calibrated relative to variations in real-world behavior (e.g., missed days of work, suicide attempts, hours of nightly sleep), then the threshold for identifying whether or not a client is diagnosis free is arbitrary. Because outcome measures employed in clinical science generally suffer from the limitation of arbitrary metrics, it is quite difficult to definitively discern whether outcome measures differ in their reliably identifying important intervention effects.

Presently, the RPC Model is agnostic as to whether some measures are better than others in identifying instances in which interventions definitively make a difference in the lives of those being treated. Thus, the framework treats individual outcome measures equally in identifying consistencies in intervention effects. Advances in model development for identifying EBIs would benefit from gauging consistencies in intervention effects in a way that weights measures in terms of whether they reflect

changes in important real-world behaviors indicative of the behaviors targeted for intervention. Future theoretical work on identifying EBIs ought to address this important domain of intervention research.

Third, criteria regarding content and number of outcome measures excluded a number of studies that might have been included in reviews employing other approaches (e.g., categorical classification criteria, traditional quantitative review methods). This suggests that the RPC Model might be limited in whether it can be applied to prior studies of interventions that were themselves limited in the number of measures employed to assess intervention outcomes. For instance, adult intervention studies often rely exclusively on self-report (Achenbach, 2006). At the same time, inconsistencies are consistently observed across studies of both adults and youths, and we have illustrated and identified inconsistencies in studies that employed multiple outcome measures of the same construct. Furthermore, this study illustrated how these inconsistencies may reveal important information of the pattern of intervention effects. Therefore, the RPC Model might encourage researchers who have previously employed limited measurement strategies to expand their conceptualizations and assessments of intervention outcomes.

Research Implications

Our findings have important implications for research and theory on intervention effects. Specifically, the multiple outcomes clinical scientists employ to evaluate intervention effects systematically vary in how they assess and estimate intervention effects. Stated another way, methods of evaluating interventions vary by measurement source, measurement method, and method of statistical analysis, and the findings gleaned from these methods exhibit identifiable patterns of effects. Prior work has viewed these inconsistencies as a hindrance to drawing research conclusions, and indeed such observations have often been seen as synonymous with identifying evidence supportive of the ineffectiveness of these interventions (e.g., De Los Reyes & Kazdin, 2006, 2008; Westen & Morrison, 2001). We disagree with this view of inconsistencies for two reasons. First, there are no definitive methods for determining whether patterns of findings indicate that some measures are "right" and others "wrong." Second, recent work suggests that measurements of inconsistencies in reports of the same behaviors (e.g., parent vs. teacher reports of children's behavior; parent and child reports of each other's behavior) predict outcomes longitudinally and relate to variations in behavior observed in the laboratory (Beck, Hartos, & Simons-Morton,

2006; De Los Reyes, Henry, Tolan, & Wakschlag, 2009; Ferdinand, van der Ende, & Verhulst, 2004; Guion, Mrug, & Windle, 2009; Pelton, Steele, Chance, & Forehand, 2001). That is, measurements of inconsistencies between parallel reports of psychosocial behaviors demonstrate adequate psychometric properties (De Los Reyes, Goodman, Kliewer, & Reid-Quiñones, 2008; De Los Reyes & Kazdin, 2004). Consistent with this work, we found that using the RPC Model researchers can employ previously underutilized outcomes patterns as a new resource for discovering the circumstances in which interventions yield robust effects.

An example here may be helpful. For instance, for BPT studies our findings revealed that in the majority of circumstances in which consistent effects were identified, these effects were specific to parent-reported outcomes. This specificity might suggest ways in which an intervention exacts particular effects and how an intervention might be modified to enhance these effects. Thus, in a study of a parent-focused intervention for childhood conduct problems, consistent effects among groups of outcomes based on parent report and inconsistent effects among groups of teacher-reported outcomes might suggest that the intervention works particularly well in changing contingencies influencing the targeted behavior when expressed in home settings (e.g., deviant parent–child interactions; parent's poor behavior management strategies). This interpretation is supported by recent laboratory evidence suggesting that parent and teacher reports of disruptive behavior in young children, and more particularly, the discrepancies between their reports, map onto differences in observations of disruptive behavior in these same children, depending on the adult with whom they are interacting (De Los Reyes et al., 2009). In this study, when only the parent identified disruptive behavior, this report related to observed disruptive behavior within parent–child interactions and not clinical examiner–child interactions. Conversely, when only the teacher identified disruptive behavior, this report related to observed disruptive behavior within examiner–child interactions and not parent–child interactions. Thus, identifying inconsistent outcomes might not suggest weaker intervention effects, but rather different intervention effects, because outcomes might vary depending on the context(s) in which interventions exact effects and/or the context(s) in which the targeted behavior is primarily expressed.

Acknowledging and examining inconsistencies among outcomes might result in a greater sense of how to administer that intervention in the future and what groups of outcomes might be expected to suggest consistent effects. Indeed, this knowledge might lead to future work typified by improved specification of variations in behavior, and revisions to the

intervention that specifically target this behavioral variation. For instance and in relation to the example above, a controlled trial might be conducted to experimentally test the parent-reported specificity of the intervention's effects. Specifically, one could examine whether targeting behaviors in a home-specific manner (intervention techniques specifically targeting dysfunctional parent–child interactions and/or parents' management skills) yields more pervasive effects than a situation-nonspecific version of the intervention (intervention techniques generally targeting multiple adult–child interactions [parents, teachers, unfamiliar adults] and interactions with the child and his or her peers). Therefore, inconsistencies yield fruitful knowledge of intervention effects because they can guide hypothesis generation and research design. We encourage future research to examine whether patterns of findings in outcome studies relate to situation-specific variations in clinical change.

Clinical Implications

The findings from this study yield useful knowledge with implications for clinical practice. Indeed, outcome measures are often based on reports taken from those involved in the treatments being tested (e.g., clinicians, parents, children). Thus, those involved in the provision of psychological services may differ greatly in whether they perceive successful intervention outcomes. At the same time, our findings suggest that one can identify patterns of significant effects within these differences. As a result, the findings reported in controlled trials might suggest how the intervention may work in clinic settings. Therefore, when clinicians look to the intervention literature to inform their practice, selections of interventions might be guided by two critical considerations: (1) What kinds of benefits were observed in research (e.g., which informants consistently observed change, what kinds of changes were observed)? and (2) do these benefits fit the nature and extent of the targeted behaviors or clinical presentation of the client? For instance, if the referral problem for a child client was aggressive behavior that was primarily impairing school performance, an intervention might not be useful for them if the effects reported in the literature were specific to informants who often do not have direct access to observations of the child's behavior in school settings (e.g., parents). However, if the primary foci of the intervention were dysfunctional parent–child interactions related to the child's oppositional behavior in the home, then the intervention might yield beneficial changes that consumers of the intervention can relate to and view as particularly helpful.

If practitioners focus on the patterns of consistencies in evidence suggested by studies, the likelihood of consumers' perceived effectiveness of interventions increases substantially. In turn, this makes research an exponentially potent information source for clinical practice and the selection of intervention techniques. Furthermore, research-reporting guidelines might benefit from advising researchers to report a summary of results in a table format at the end of studies that reveals where consistencies were observed and not observed. In sum, greater clarity in reporting patterns of findings in clinical trials will increase the likelihood of integration of research into practice and provide practitioners with tools for enhancing the utility of research in informing the provision of health care services.

Notes

1. An expanded version of the Method for this article providing further details on such domains as the rationale for examining CBT and BPT interventions, study inclusion criteria, information on excluded studies, and study coder training is available from the authors by request or online at: http://sites.google.com/site/caipumaryland/Home/about-caip/caip-lab-publications

2. By "employed," we mean that studies must have prospectively administered at least three outcome measures of the target construct (i.e., pre- and post-intervention), and sufficient data must have been reported in the published study to calculate effect sizes and tests of statistical significance. By "three outcome measures," we mean that the authors must have employed three measures that were each distinctly administered from each other (e.g., a study was not included if the authors only reported three findings from subscale scores gleaned from the same measure).

3. Intervention-control comparisons were coded under a single effect size range, even if they were ascribed more than one RPC Model category. For example, if an intervention-control comparison yielded evidence specific to parent-rated outcomes, measured via questionnaire, then the effect size range for this comparison would encompass only the findings within these category classifications. Similarly, if an intervention-control comparison was classified under an RPC Model category denoting specificity of change (Table 1), then the effect size range only encompassed findings within this category classification.

References

*References marked with an asterisk indicate studies included in the meta-analysis.

Achenbach, T. M. (2006). As others see us: Clinical and research implications of cross-informant correlations for psychopathology. *Current Directions in Psychological Science, 15*, 94-98.

Achenbach, T. M., Krukowski, R. A., Dumenci, L., & Ivanova, M. Y. (2005). Assessment of adult psychopathology: Meta-analyses and implications of cross-informant correlations. *Psychological Bulletin, 131*, 361-382.

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin, 101*, 213-232.

American Psychological Association Interdivisional Task Force on Child and Adolescent Mental Health. (2007). *Links for finding evidence-based interventions.* Retrieved September 4, 2007, from http://ucoll.fdu.edu/apa/ lnksinter.html

Ball, C. N. (2003). *Georgetown linguistics: Web chi square calculator.* Retrieved January 22, 2007, from http://www.georgetown.edu/faculty/ballc/webtools/web_chi.html

*Barrett, P. M., Dadds, M. R., & Rapee, R. M. (1996). Family treatment of childhood anxiety: A controlled trial. *Journal of Consulting and Clinical Psychology, 64*, 333-342.

Beck, K. H., Hartos, J. L., & Simons-Morton, B. G. (2006). Relation of parent-teen agreement on restrictions to teen risky driving over 9 months. *American Journal of Health Behavior, 30*, 533-543.

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist, 61*, 27-41.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

De Los Reyes, A., Henry, D. B., Tolan, P. H., & Wakschlag, L. S. (2009). Linking informant discrepancies to observed variations in young children's disruptive behavior. *Journal of Abnormal Child Psychology, 37,* 637-652.

De Los Reyes, A., Goodman, K. L., Kliewer, W., & Reid-Quiñones, K. R. (2008). Whose depression relates to discrepancies? Testing relations between informant characteristics and informant discrepancies from both informants' perspectives. *Psychological Assessment, 20*, 139-149.

De Los Reyes, A., & Kazdin, A.E. (2004). Measuring informant discrepancies in clinical child research. *Psychological Assessment, 16*, 330-334.

De Los Reyes, A., & Kazdin, A. E. (2005). Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin, 131*, 483-509.

De Los Reyes, A., & Kazdin, A. E. (2006). Conceptualizing changes in behavior in intervention research: The range of possible changes model. *Psychological Review, 113*, 554-583.

De Los Reyes, A., & Kazdin, A. E. (2008). When the evidence says, "Yes, no, and maybe so": Attending to and interpreting inconsistent findings among evidence-based interventions. *Current Directions in Psychological Science, 17*, 47-51.

Ferdinand, R. F., van der Ende, J., & Verhulst, F. C. (2004). Parent-adolescent disagreement regarding psychopathology in adolescents from the general population as a risk factor for adverse outcome. *Journal of Abnormal Psychology, 113*, 198-206.

*Flannery-Schroeder, E. C., & Kendall, P. C. (2000). Group and individual cognitive-behavioral treatments for youth with anxiety disorders: A randomized clinical trial. *Cognitive Therapy and Research, 24*, 251-278.

*Gallagher, H. M., Rabian, B. A., & McCloskey, M. S. (2004). A brief group cognitive-behavioral intervention for social phobia in childhood. *Journal of Anxiety Disorders, 18*, 459-479.

Graphpad Software, Inc. (2005). *Quickcalcs online calculators for scientists: t Test calculator.* Retrieved January 22, 2007, from http://www.graphpad.com/quickcalcs/ttest1.cfm?Format=SD

Guion, K., Mrug, S., & Windle, M. (2009). Predictive value of informant discrepancies in reports of parenting: Relations to early adolescents' adjustment. *Journal of Abnormal Child Psychology, 37*, 17-30.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* San Diego, CA: Academic Press.

James, A., Soler, A., & Weatherall, R. (2005). Cognitive behavioural therapy for anxiety disorders in children and adolescents. *Cochrane Database of Systematic Reviews*. New York: John Wiley.

Kazdin, A. E. (2006). Arbitrary metrics: Implications for identifying evidence-based treatments. *American Psychologist, 61*, 42-49.

*Kendall, P. C. (1994). Treating anxiety disorders in children: Results of a randomized clinical trial. *Journal of Consulting and Clinical Psychology, 62*, 100-110.

*Kendall, P. C., Flannery-Schroeder, E., Panichelli-Mindel, S. M., Southam-Gerow, M., Henin, A., & Warman, M. (1997). Therapy for youths with anxiety disorders: A second randomized clinical trial. *Journal of Consulting and Clinical Psychology, 65*, 366-380.

*King, N. J., Tonge, B. J., Mullen, P., Myerson, N., Heyne, D., Rollings, S., et al. (2000). Treating sexually abused children with posttraumatic stress symptoms: A randomized clinical trial. *Journal of the American Academy of Child and Adolescent Psychiatry, 39,* 1347-1355.

Koenig, K., De Los Reyes, A., Cicchetti, D., Scahill, L., & Klin, A. (2009). Group intervention to promote social skills in school-age children with pervasive developmental disorders: Reconsidering efficacy. *Journal of Autism and Developmental Disorders, 39,* 1163-1172.

*Leal, L. L., Baxter, E. G., Martin, J., & Marx, R. W. (1981). Cognitive modification and systematic desensitization with test anxious high school students. *Journal of Counseling Psychology, 28*, 525-528.

*Leung, C., Sanders, M. R., Leung, S., Mak, R., & Lau, J. (2003). An outcome evaluation of the implementation of the Triple P-Positive Parenting Program in Hong Kong. *Family Process, 42*, 531-544.

Lonigan, C. J., Elbert, J. C., & Johnson, S. B. (1998). Empirically supported psychological interventions for children: An overview. *Journal of Clinical Child Psychology, 27*, 138-145.

*McMurray, N. E., Bell, R. J., Fusillo, A. D., Morgan, M., & Wright, F. A. C. (1986). Relationship between locus of control and effects of coping strategies on dental stress in children. *Child & Family Behavior Therapy, 8*, 1-17.

Nathan, P. E., & Gorman, J. M. (Eds.) (2007). *A guide to treatments that work* (3rd ed.). New York: Oxford University Press.

Pelton, J., Steele, R. G., Chance, M. W., & Forehand, R. (2001). Discrepancy between mother and child perceptions of their relationship: II. Consequences for children considered within the context of maternal physical illness. *Journal of Family Violence, 16*, 17-35.

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology, 52*, 59-82.

Roth, A., & Fonagy, P. (2005). *What works for whom?: A critical review of psychotherapy research* (2nd ed.). New York: Guilford.

*Spence, S. H., Donovan, C., & Brechman-Toussaint, M. (2000). The treatment of childhood social phobia: The effectiveness of a social skills training-based, cognitive-behavioral intervention, with and without parental involvement. *Journal of Child Psychology and Psychiatry, 41*, 713-726.

*Webster-Stratton, C. (1984). Randomized trial of two parent-training programs for families with conduct-disordered children. *Journal of Consulting and Clinical Psychology, 52*, 666-678.

*Webster-Stratton, C. (1990). Enhancing the effectiveness of self-administered videotape parent training for families with conduct-problem children. *Journal of Abnormal Child Psychology, 18*, 479-492.

*Webster-Stratton, C. (1992). Individually administered videotape parent training: "Who benefits?" *Cognitive Therapy and Research, 16*, 31-52.

*Webster-Stratton, C., & Hammond, M. (1997). Treating children with early-onset conduct problems: A comparison of child and parent training interventions. *Journal of Consulting and Clinical Psychology, 65*, 93-109.

*Webster-Stratton, C., Kolpacoff, M., & Hollinsworth, T. (1988). Self-administered videotape therapy for families with conduct-problem children: Comparison with two cost-effective treatments and a control group. *Journal of Consulting and Clinical Psychology, 56*, 558-566.

*Webster-Stratton, C., Reid, M. J., & Hammond, M. (2004). Treating children with early-onset conduct problems: Intervention outcomes for parent, child, and teacher training. *Journal of Clinical Child and Adolescent Psychology, 33*, 105-124.

Weisz, J. R., Hawley, K. M., & Jensen Doss, A. (2004). Empirically tested psychotherapies for youth internalizing and externalizing problems and disorders. *Child and Adolescent Psychiatric Clinics of North America, 13*, 729-815.

Weisz, J. R., Jensen Doss, A., & Hawley, K. M. (2005). Youth psychotherapy outcome research: A review and critique of the evidence base. *Annual Review of Psychology, 56*, 337-363.

Weisz, J. R., Weiss, B., Alicke, M. D., & Klotz, M. L. (1987). Effectiveness of psychotherapy with children and adolescents: A meta-analysis for clinicians. *Journal of Consulting and Clinical Psychology, 55*, 542-549.

Weisz, J. R., Weiss, B., Han, S. S., Granger, D. A., & Morton, T. (1995). Effects of psychotherapy with children and adolescents revisited: A meta-analysis of treatment outcome studies. *Psychological Bulletin, 117*, 450-468.

Weisz, J. R., McCarty, C. A., & Valeri, S. M. (2006). Effects of psychotherapy for depression in children and adolescents: A meta-analysis. *Psychological Bulletin, 132*, 132-149.

Westen, D., & Morrison, K. (2001). A multidimensional meta-analysis of treatments for depression, panic, and generalized anxiety disorder: An empirical examination of the status of empirically supported therapies. *Journal of Consulting and Clinical Psychology, 69*, 875-899.

Woolfenden, S. R., & Williams, K., & Peat, J. (2001). Family and parenting interventions in children and adolescents with conduct disorder and delinquency aged 10-17. *Cochrane Database of Systematic Reviews*. New York: John Wiley.

Andres De Los Reyes is an assistant professor of psychology at the University of Maryland-College Park and director of the Comprehensive Assessment and Intervention Program. He received his PhD in clinical psychology from Yale University and serves on the Editorial Board of *Journal of Clinical Child and Adolescent Psychology.*

Alan E. Kazdin is the John M. Musser Professor of Psychology and Child Psychiatry at Yale University and director of the Yale Parenting Center and Child Conduct Clinic. He received his PhD in clinical psychology from Northwestern University and in 2008 he was president of the American Psychological Association.